Privacy Preserving Cyberbullying Prevention with AI Methods in 5G Networks

Ramezanian, Sara

FRUCT Oy

2019

Ramezanian , S & Niemi , V 2019 , Privacy Preserving Cyberbullying Prevention with AI Methods in 5G Networks . in S Balandin , V Niemi & T Tuytina (eds) , Proceedings of the 25th Conference of Open Innovations Association FRUCT, Helsinki, Finland . Proceedings of the ... Conference of Open Innovations Association FRUCT , FRUCT Oy , Helsinki , pp. 265-271 , Proceedings of the 25th Conference of Open Innovations Association FRUCT , Helsinki , Finland , 05/11/2019 . https://doi.org/10.23919/fruct48121.2019.8981521

http://hdl.handle.net/10138/309997
https://doi.org/10.23919/fruct48121.2019.8981521

Privacy Preserving Cyberbullying Prevention with AI Methods in 5G Networks

Sara Ramezanian, Valtteri Niemi
University of Helsinki and
Helsinki Institute for Information Technology
Helsinki, Finland
sara.ramezanian, valtteri.niemi@helsinki.fi

***Abstract*—Children and teenagers that have been a victim of bullying can possibly suffer its psychological effects for a lifetime. With the increase of online social media, cyberbullying incidents have been increased as well. In this paper we discuss how we can detect cyberbullying with AI techniques, using *term frequency-inverse document frequency*. We label messages as benign or bully. We want our method of cyberbullying detection to be privacy-preserving, such that the subscribers' benign messages should not be revealed to the operator. Moreover, the operator labels subscribers as normal, bully and victim. The operator utilizes policy control in 5G networks, to protect victims of cyberbullying from harmful traffic.**

I. Introduction

With widespread of Internet, online social interactions are becoming more and more common. Statistics show that the number of social media users are increasing worldwide [1], and therefore, the impact of social media on children, teens, adults, and families are undeniable [2].

Detecting the presence of cyberbullying in an online conversation, text messages and images, has been an interesting topic for researchers and there is significant number of studies in this regards, such as [3] and [4].

Commonly, cyberbullying is referred to repeatedly and intentionally posting an aggressive context via an electronic medium such as social media, text messages, emails, blogs, etc. [5]. Studies show that cyberbullying can have permanent effects on the self-esteem of the involved parties (the victim and the offender) [6]. This shows the importance of detecting and preventing cyberbullying.

Although, there have been several studies on how to detect cyberbullying with the help of Artificial Intelligence (AI) methods [7], [8], it is still difficult to instantly detect a harmful electronic context [7], and protect victim's of cyberbullying to receive such context. On the other hand, individual's messages and online comments are considered to be private. Therefore, it makes sense to carry out the detection of cyberbullying in a privacy-preserving manner.

With the emerging of 5th generation of mobile networks (5G networks), data availability and information process will be even faster and easier than before. One of the key functionality of 5G networks is that we can add additional functions, to the architecture of the system. In this paper, we investigate the following research questions: Could cyberbullying protection be carried out in the 5G networks, as a new function? Can we design a cyberbullying detection method that is privacy-preserving, and can also protect the victims from being exposed to such harmful contents? Moreover, we want to investigate how we can use AI techniques as part of this method.

Our contribution in this work consists of the paradigm of a privacy-preserving cyberbullying prevention system that utilizes Artificial Intelligence techniques, and is applicable in 5G mobile networks. Another contribution is description of the solution in architectural level. The details of specifications and the implementation of our solution are left for future work.

The rest of this paper is structured as follows. Section II gives the preliminaries and required concepts that are needed to follow the rest of the paper. In Section III, we present the problem statement. Then, in Section IV, we explain the state of the art in cyberbullying detection. We detail our privacy-preserving system of cyberbullying prevention in Section V. Also, we explain where our prevention happens in the architecture of 5G networks. Finally, we conclude the paper in Section VI, and give some directions to future work.

II. PRELIMINARIES

In order to privately detect cyberbully incidents, we use several privacy-preserving techniques and AI methods. Moreover, we explain how our model is applicable to 5G networks. Therefore, in this section, we present the preliminary concepts that are required to understand the rest of this paper in three subsections: AI methods, 5G networks, and Privacy-preserving techniques.

A. AI Methods

The term *Artificial Intelligence* (AI), refers to the act of intelligence (learning and problem solving) that can be demonstrated by machines [9].

Here, we present the concepts of AI methods that we use to design our privacy-preserving cyberbullying prevention system.

1) Term Frequency-Inverse Document Frequency (TF-IDF): In order to reflect the importance of a word or phrase to the whole document corpus, a numerical statistic called tf-idf, can be used [10]. We assume that a document consists of several comments. The tf-idf score is computed as follows:

$$\text{tf-idf} = tf_{ij} * idf_i. \quad (1)$$

In equation (1), i is a word or phrase in comment j. The term frequency tf_{ij} is a score that determines the importance of a word i in j, and calculated as follows:

$$tf_{ij} = \frac{F_{ij}}{\sum F_j} \quad (2)$$

where F_{ij} is a score to determine how frequent is a word i in a comment j. Moreover, the frequency of all the words in a comment j is shown by $\sum F_j$.

To compute how important a word i is to the entire corpus of the document, denoted as idf_i, we use the following equation:

$$idf_i = \frac{\log n(C)}{n(i \in C)}. \quad (3)$$

In equation (3), the number of comments is denoted by $n(C)$, and $n(i \in C)$ is the number of comments in which the word i appeared.

The above calculations reveal the relevance of a certain word to the entire document [11]. Therefore, it is a useful tool to detect an intentional act of harassment that has happened in a text message.

2) F1-Score: In order to measure the accuracy of a test that has been done with a binary classifier,a statistic score is used that is called F1-score [12].

F1-score is the harmonic mean [13] of precision p and recall r:

$$\text{F1-score} = 2.\frac{pr}{p+r}. \quad (4)$$

Precision is determined by:

$$\text{precision} = \frac{\text{relevant retrieved instances}}{\text{total amount of retrieved instances}} \quad (5)$$

while recall is evaluated as:

$$\text{recall} = \frac{\text{relevant retrieved instances}}{\text{total amount of relevant instances}}. \quad (6)$$

B. Privacy Preserving Techniques

In this part, the techniques that makes our cyberbullying prevention method to function in a privacy-preserving manner, have been presented.

1) Cryptographic Hash Function: A one-way function that maps input of arbitrary size to a bit string (hash value) that has a fixed size, is called a cryptographic hash function (hereafter *hash function*) [14].

The properties of a secure hash function $h(x)$ are as follows:

- Collision resistance: It is infeasible to find two input x and x' such that $x \neq x'$, and $h(x) = h(x')$.
- Preimage-resistance: For any output y, it is infeasible to find an input x', such that $h(x') = y$.
- Deterministic: The same input always results to the same hash value.
- It is fast and easy to compute.

2) Private Set Intersection (PSI): A PSI protocol is a cryptographic protocol that is executed between two parties. The input of each party is a private set, and the output of the protocol is the intersection of the two sets, without revealing any information about the elements in the sets that are not in the intersection [15].

C. 5G Networks

5G is the next generation of the mobile telecommunication network that evolves from 4G/LTE system. Traditionally, cellphones were the only subscribers of mobile networks. 5G systems as an immersive connectivity technology cover a broad range of new use cases such as Internet of Things (IoT), Virtual Reality/Augmented Reality (VR/AR), etc. To meet the requirements of these versatile verticals, 5G aims to offer some key goals such as very high throughput (1-20 Gbps), ultra-low latency (<1ms), massive connectivity, and low energy consumption, to name a few.

The reference point architecture of the 5G system [16] is as follows. Users connect to the network through radio stations, called New Generation Node Base station (gNB). All the other components involved in the network, except the User Equipment (UE) and Radio Access Network (RAN), form the so-called *core network*. To increase flexibility, and realize new concepts, the core network itself has been separated into two logical plane functions: User Plane Function (UPF), and Control Plane Function (CPF). Fig. 1 shows 5G system from a Service Based Architecture (SBA) view.

In 5G SBA, each control function initially registers its service to Network Repository Function (NRF) and later finds other functions' services through a service discovery request against NRF. Network functions communicate with other authorized functions through a set of standard APIs. The complete list of network functions and their functionality [16], [17] is beyond the scope of this paper. In the following, we only introduce a couple of functions that will be used later in our proposed solution.

1) User Plane Function (UPF): The UPF is the principle function for handling the data traffic from UE which runs tasks such as packet routing and forwarding, packet inspection and QoS handling.

2) Session Management Function (SMF): The SMF selects and controls UPF for traffic routing, among other services. It acts as the interface for all communication related to offered user plane services. SMF determines how the policy and charging for these services are applied.

3) User Data Management (UDM): The UDM acts as the main function to store subscription information. Among others, it stores the long-term security credentials used in authentication, subscription identifiers, and access authorization data.

4) Policy Control Function (PCF): The PCF is the corresponding function to PCRF in 4G/LTE system. The function accesses the User Data Repository (inside UDM) to collect subscription information for policy decisions. It provides policy rules to be used by other control plane functions.

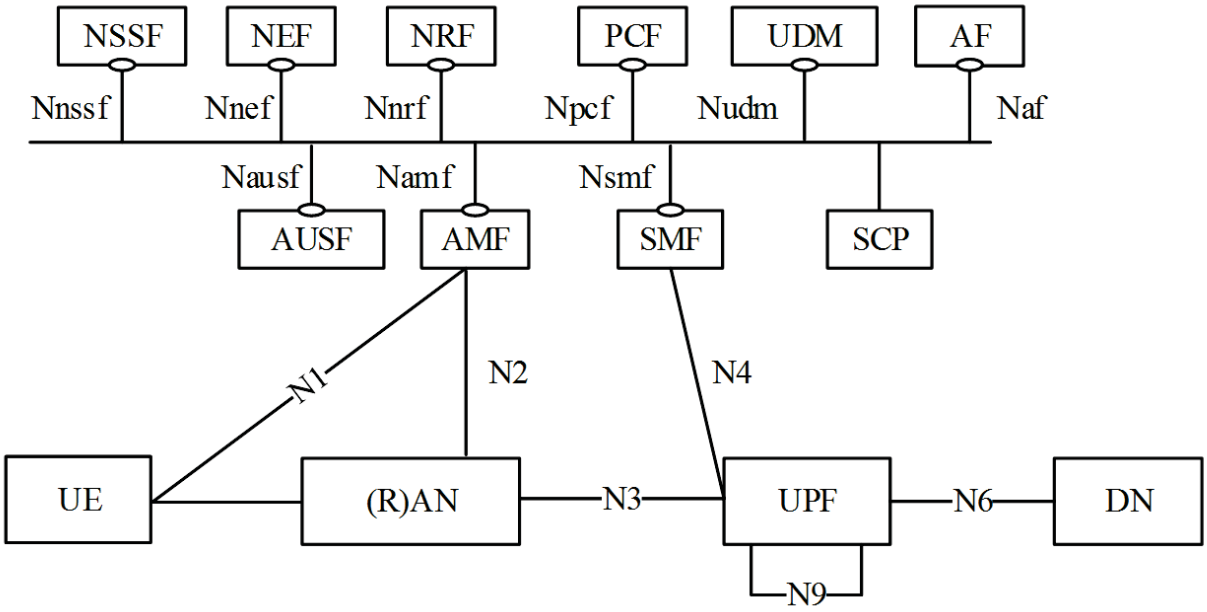

Fig. 1. 5G system from a Service Based Architecture (SBA) view. Source: Fig. 4.2.3-1 in [16]

III. PROBLEM STATEMENT

As we explained before, cyberbullying can have a lifetime bad psychological effects on its victims as well as on the offenders [18], [19]. Moreover, automatic detection of a cyberbullying incident is difficult because the nature of bullying is subjective [8].

On the other hand, typically cyberbullying detection takes place only after the victim has been exposed to the damaging context [7].

Studies on cyberbullying detection are mostly not considering the privacy of the sender nor the receiver of the message, and directly analyze the context of the message.

With the emerge of 5G networks, accessing to data will be even faster than before, and this may cause to a bigger venue for bullies to find and harass their victims. Moreover, 5G promises to cover more human activities, therefore, it can offer cyberbullying prevention. In this paper, we study methods to design a privacy-preserving cyberbullying prevention system, in 5G networks.

IV. RELATED WORK

To our best knowledge, there has not been any work on privacy preserving cyberbullying detection in 5G networks.

Here, we present the related work on cyberbullying detection with AI techniques. We categorize the detection techniques as *Content Based* and *Other Techniques*. Content based cyberbullying detection techniques are categorized as *Direct* and *Indirect*.

A. Content Based

In order to detect a cyberbullying incident from a text, some studies design classifiers with content based features, such as cyberbullying keywords, term frequency-inverse document frequency, etc. Here, we present some of those studies.

1) Direct : The presence of cyberbullying keywords (threat, curse, racism, hate and sexism) in online texts is the easiest way to detect a cyberbullying incident, because such message is intended to harm its receiver, and therefore, most probably contains harassment and abusive keywords [8], [20].

Features such as *number of profanity words* and *ratio of capital letters* to detect shouting [21], *use of pronoun 'you'* [10], *pattern of harassment* such as 'you + bullying keyword', and *length of the comment* [22], *usernames containing bad words* [23], *the percentage of curse and insult words* [24], are some of the direct content based features in cyberbullying detection classifiers.

2) Indirect : Sometimes, cyberbullying happens without using any direct profanity words, or aggressive ones, in a message. Spreading False rumors and telling lies about victim, tagging the victim's embarrassing images, refuse to socialize with victim (e.g. not replying to their messages), repetition of a bad nickname, etc. are examples of bullying that does not contain the cyberbullying keywords [25], [26].

B. Other Techniques

Here, we present some of the other features that are used to detect cyberbullying, that do not depend on context of the message.

One such approach is to enhance the techniques by analyzing social network features [27]. Taking gender specific language features, age and gender, can also be effective to determine cyberbullying incidents [28].

For social media accounts, detecting fake and suspicious profiles, number of uploads, subscriptions, comments are

TABLE I. TABLE OF NOTATIONS

Definition	Notation
The message consists of l words	$m = w_1 + w_2 + ... + w_l$
Database of k cyberbullying keywords	$B = \{b_1, b_2, ..., b_k\}$
The set of time stamps	t
point of bully	p_b
point of victim	p_v
Threshold for bully	τ_b
Threshold for victim	τ_v

features that are used in cyberbullying detection classifiers [23].

V. OUR METHOD

In this section, we explain in details how to design a privacy-preserving cyberbullying prevention system in 5G networks, that utilizes AI techniques.

We assume that the operators offer different services to customers, as adults and children. In our method, we prevent children from online harassment and cyberbullying. However, our method can be generalized and be used against cyberbullying in general.

Our method is designed for messaging services that the operator provides (such as SMS and MMS), and other messaging applications that have their own services inside the network.

Our method has three phases. In phase 1, we explain how to detect whether a text message contains cyberbullying keywords, in a privacy-preserving way. Phase 2, comes to play only when the results of phase 1 is positive. In phase 2, with the help of AI methods, we classify the message as either bully or benign. Phase 3, is where the operator decides how to label the subscribers in order to define policy controls according to each subscriber's label.

We use different parameters in our system and here, we list some of them. The set of time stamps is denoted by t. We define two parameters p_b and p_v as point of bully and point of victim, respectively. τ_b and τ_v are notations to show threshold for bully and threshold for victim, respectively. Table 1, shows the parameters that we use in our system.

A. Phase 1

To preserve the privacy of the subscribers, we aim to design a method to detect cyberbullying instances without revealing the context of a benign message to the operator. To this regard, we create a *Filter Check* phase in our system, to flag messages that potentially contain harassment or bullying context.

We assume that the sender has a message m that consists of l words, such that $m = w_1 + w_2 + ... + w_l$, and the operator has a set of cyberbullying keywords $B = \{b_1, b_2, ..., b_k\}$ that contains k. words.

Here, we present in detail three privacy preserving protocols to detect cyberbullying. In all three protocols, we assume that the parties are honest and follow the protocol as explained here. The protocols are presented in order of the most secure to the least secure one. At the same time, they are presented in order of the least efficient to most efficient.

In Fig. 3, phase 1 is presented as Filter Check.

1) Using Private Set Intersection: The sender has a message m which is a set of l words, and the operator has a set B consists of k cyberbullying keywords. By utilizing any private set intersection protocol, the sender and the operator can find the intersection between their two sets, such that the context of the message remains private. After executing the protocol only if the message contains any of those cyberbullying keywords, the private set intersection protocol results in positive, and otherwise, results in negative.

Now, we explain a private set intersection protocol that only shows whether the intersection of input sets are empty or non-empty. The sender and the operator each pick a large integer a and b, respectively. For each word w_i in the message m, the sender computes ${w_i}^a$, and sends them to the operator. For each ${w_i}^a$, the operator computes and stores ${w_i}^{ab}$. Then, for each b_j in the database of bullying-keywords, the operator computes ${b_j}^b$, and sends them to the sender. For each ${b_j}^b$, the sender computes ${b_j}^{ba}$, and sends them to the operator. Now, the operator finds the intersection between the set of ${w_i}^{ab}$ and the set of ${b_j}^{ba}$. An overview of this private set intersection protocol is presented in Fig V-A1.

If the intersection is non-empty, the operator asks for the plaintext m, and follows phase 2 to check whether m is actually contains the traces of cyberbullying, with AI techniques. Otherwise, the operator forwards the message to its receiver.

2) Using Hash Functions: The sender and the operator together pick a hash function H. The sender has a message $m = w_1 + w_2 + ... + w_l$. In order to hide the words in the message m, the sender adds l random clean words to the message, shuffles it and creates a new message m'. Then, the sender computes the hash value of each word in the message m' and sends $H(w_1) + H(w_2) + ... + H(w_{2l})$ to the operator. The operator has a list of cyberbullying keywords and computes the hash value of each word in the list. Then, the operator computes the set intersection of its hash values with the hash values in the text m' that has been sent by the sender. If the intersection is empty, the message consider to be benign and the operator forwards the message to its receiver. Otherwise, the operator asks for the message m and use the AI techniques in phase 2, to check whether the message is actually a cyberbully.

3) Using a Trusted Party: We can utilize a trusted party as median between the sender and the operator. In this approach, the trusted party keeps the database of cyberbullying keywords. The sender sends the message m to the trusted party, where the message is checked against the database of cyberbullying keywords. If the message contains word(s) that suggest the message is a cyberbully, the trusted party marks the message as bully and sends the message to the operator. Then, the operator starts phase 2. If the message does not have any of the cyberbullying keywords, the operator will get this information together with the message. Then, the operator

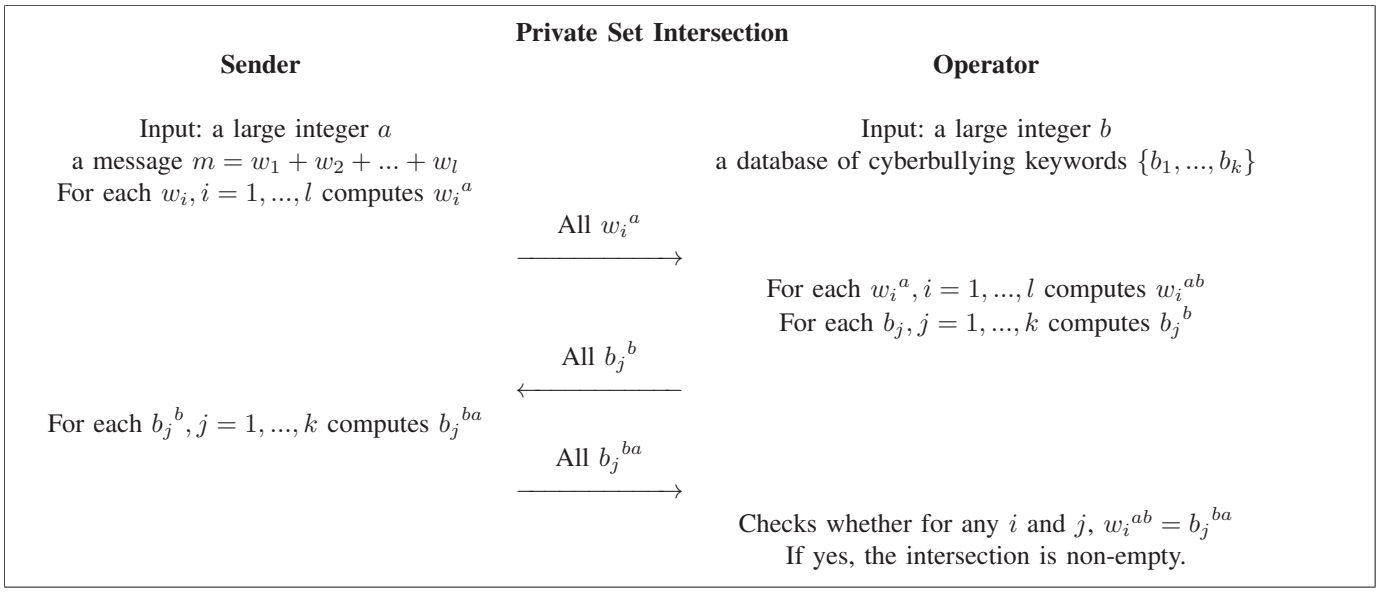

Fig. 2. An overview of a PSI protocol

forwards the message to the receiver.

B. Phase 2

After a message has been tested in the filter check (phase 1) and it has been determined that the message contains at least one of the cyberbullying keywords, the operator initiates phase 2. In this phase, the message is examined with the help of AI methods, to be classified either as cyberbully or benign.

In this work, in order to detect cyberbullying with AI techniques we use the results of a recent work by Rosa et al. [7]. We particularly use the results of Fig. 1 in [7], where the authors determined which combinations of features result in better detection of cyberbullying instances. Based on their performance analysis, which is based on F1 scores, the tf-idf approach results in better detection of cyberbullying, and adding more features, such as *personality trait, textual, word embedding etc.* features, may create noise and therefore, reduce the performance of the classifiers.

If the test result show that the message has been classified as cyberbullying, the operator marks the message as bully and otherwise as benign. Then, the operator starts phase 3.

Please note that the two phases of examinations (Phases 1 and 2), make the detection of cyberbullying incidents to be more accurate and therefore, minimize the number of false positive detections.

C. Phase 3

In this phase, we show how the privacy-preserving cyberbullying prevention takes place in 5G architecture.

The operator marks all the subscribers as normal. Each subscriber has five parameters: t, p_b, p_v, τ_b and τ_v. At first, the operator assigns value zero to p_b, p_v, τ_b and τ_v and the set t is empty.

We envision that the components described in previous subsection (phases 2) could be packaged into a single function, as it is shown in Fig. 3. We name this function as *cyberbullying prevention function* (CBPF). The service-based architecture of 5G systems makes interactions of core network functions to each other, as well as to external functions (non-3GPP applications), less complex through a unified API system.

First, the subscriber sends a message from his/her device (UE), which goes to UPF (interaction 1 in Fig. 3). In order to detect and prevent bullying subscribers in the network, the misbehaving subscribers should be flagged and continuously monitored by utilizing Filter Check (interaction 2 in Fig. 3). In our proposed solution, if subscriber's message has been flagged by the Filter Check, the message goes through SMF (interaction 3 in Fig. 3) and is then sent to CBPF (interaction 4 in Fig. 3) to check whether it contains harassment.

If the check shows traces of cyberbullying, the message should be marked as bully, and the sender's p_b and the receiver's p_v are incremented by one. The operator also marks the time and date in t. The idea here is to keep track of subscribers history in regards to cyberbullying in User Data Management (UDM) (interaction 5 in Fig. 3), e.g. accommodating some decision metrics. This subscriber profiling could be helpful as whether to forward subscriber's traffic toward some particular destinations (e.g. a victim subscriber) in later stages of the system. More concretely, the Policy Control Function (PCF) uses this information to enforce appropriate policy rules ((interaction 6 in Fig. 3)) to traffic flow in UDM, through SMF (interaction 7 in Fig. 3), in order to block harmful messages. Now, if the message is benign, it goes through and otherwise it will be blocked.

If a subscriber's p_b (p_v) reached the threshold of our system τ_b (τ_v) then he/she will be marked as bully (victim) by the operator. The parameter t can be used to report the bullying incidents, for example to parents and school's administrators.

If a message is marked as bully, the operator does not send

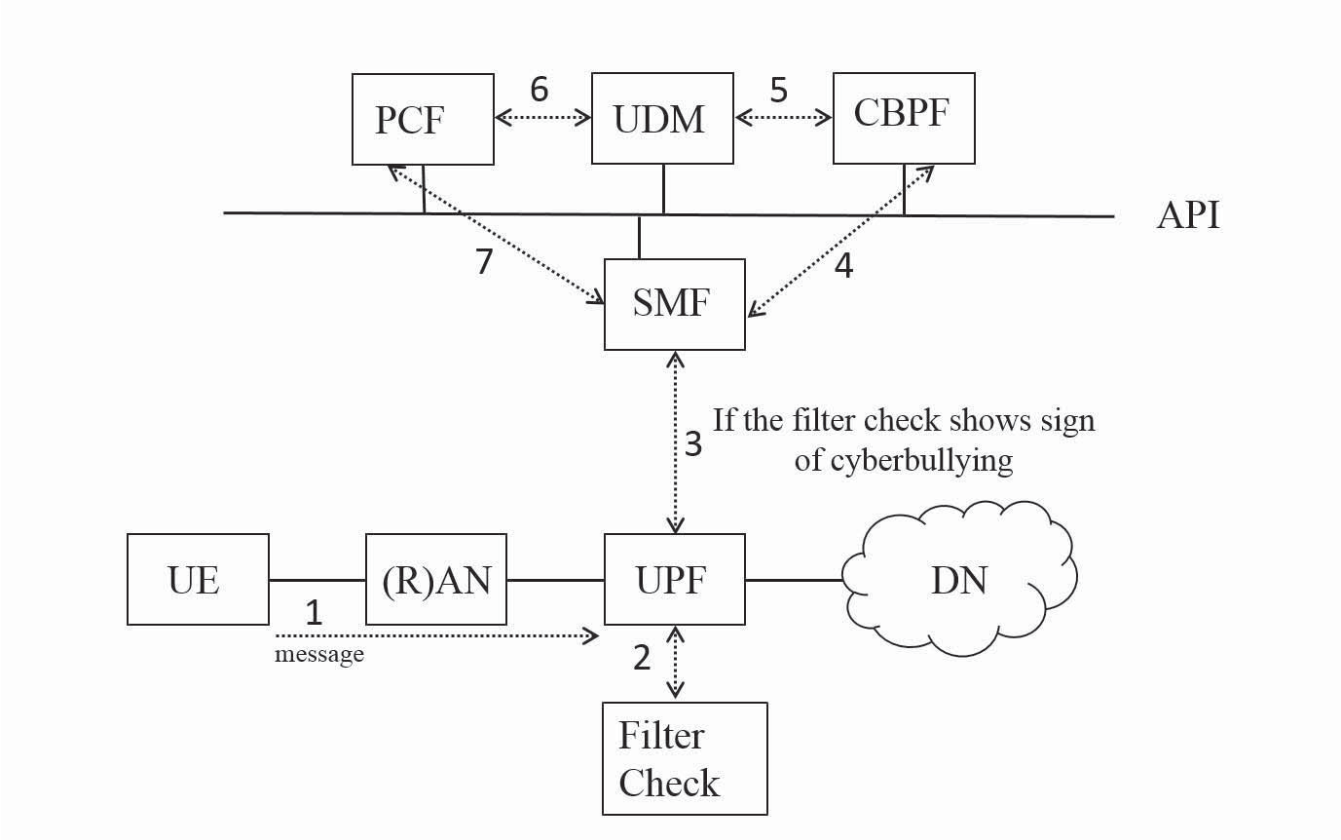

Fig. 3. A summary of our cyberbullying prevention method in 5G

the message forward. Therefore, the receiver is automatically protected from the harmful content of that message.

After some time, the subscribers will be categorized as normal, victim and bully. If a subscriber is a normal user, there is no need to Filter Check all his/her messages. Instead, only a small portion of messages will be checked for the purpose of periodic re-evaluation.

Defining the exact value for the thresholds τ_b and τ_v, requires a carefully designed user study and help of psychologists. We leave this part to the future work.

VI. Conclusion and Future Work

In this work we proposed the paradigm of our privacy-preserving cyberbullying prevention system in 5G networks that utilizes AI methods.

At first, all the subscribers are labeled as normal, by the operator. After a text message is marked as bully, the operator add one point to the value of p_b for the user who sent the message, and one point to the value of p_v for the receiver. If the value p_b (p_v) is bigger than threshold τ_b (τ_v), then the operator changes the label of that subscriber to bully (victim) and applies policy controls accordingly. In this way, a bullying message is recognized imminently and the victim will be protected automatically.

Our method is limited to harassment and bullying detection. Future work is needed to improve the system such that sarcastic messages can also be detected and if needed, consider to be bullying. For example, the sentence "I love your BIG NOSE!!", is probably intended to be harmful, however, as it does not contain any cyberbullying keywords, can not be detected by our system.

Another direction for future work is to design a privacy-preserving classifiers. Theoretically, we can pick a hash function H, and compute the hash value of all the words in the database of bullying-keywords. Then, train the machine learning model with these hash values of the keywords. In order to check a message m for traces of cyberbullying, we should first compute the hash value of each word in m, then check theses hash value with the classifier to see whether the message intends to bully. In this way, the actual message remains private, and the system will only classifies the messages as bully or benign.

Acknowledgment

This work was supported in part by 5GFORCE project, funded by Business Finland.

References

[1] Statista, Social Media Statistics & Facts, Web: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users. html.

[2] G. S. O'Keeffe, K. Clarke-Pearson, "The impact of social media on children, adolescents, and families", *Pediatrics*, 2011 , vol. 127(4), pp. 800-804.

[3] J. Bayzick, A. Kontostathis, L. Edwards, "Detecting the presence of cyberbullying using computer software", *In Ursinus College*, 2011, pp. 93-96.

[4] H. Hosseinmardi, S.A. Mattson, R.I. Rafiq, R. han, Q. Lv, S. Mishra, "Analyzing labeled cyberbullying incidents on the instagram social network", *In International conference on social informatics*, 2015, December, Springer, Cham, pp. 49-66.

[5] R.M. Kowalski, G.W. Giumetti, A.N. Schroeder, H.H. Reese, "Chapter 14 Cyber bullying among college students: Evidence from multiple domains of college life", *In Misbehavior online in higher education*, 2012, Emerald Group Publishing Limited, pp. 293-321.

[6] J.W. Patchin, S. Hinduja, "Cyberbullying and selfesteem", *In Journal of school health*, 2010, 80(12), pp.614-621.

[7] H. Rosa, N. Pereira, R. Ribeiro, P.C. Ferreira, J.P. Carvalho, S. Oliveira, L. Coheur, P. Paulino, A.V. Simo, I. Trancoso, "Automatic cyberbullying detection: A systematic review", *In Computers in Human Behavior*, 2019, 93, pp. 333-345.

[8] S. Salawu, Y. He, J. Lumsden, "Approaches to automated detection of cyberbullying: A survey", *In IEEE Transactions on Affective Computing*, 2017.

[9] S.J. Russell, P. Norvig, "Artificial intelligence: a modern approach", *In Pearson Education Limited*, 2016, Malaysia.

[10] V.S. Chavan, S.S. Shylaja, "Machine learning approach for detection of cyber-aggressive comments by peers on social media network", *In International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2015, pp. 2354-2358, IEEE.

[11] J. Ramos, "Using tf-idf to determine word relevance in document queries", *In Proceedings of the first instructional conference on machine learning*, 2003, December, Vol. 242, pp. 133-142.

[12] C. Goutte, E. Gaussier, "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation", *In European Conference on Information Retrieval*, 2005, March, pp. 345-35. Springer, Berlin, Heidelberg.

[13] W.J. Dixon, F.J. Massey Jr, "Introduction to statistical analysis", 1951, McGraw-Hill.

[14] P. Rogaway, T. Shrimpton, "Cryptographic hash-function basics: Definitions, implications, and separations for preimage resistance, second-preimage resistance, and collision resistance.", *In International workshop on fast software encryption*, 2004, February, pp. 371-388, Springer, Berlin, Heidelberg.

[15] M.J. Freedman, K. Nissim, B. Pinkas, "Efficient private matching and set intersection. In International conference on the theory and applications of cryptographic techniques.", *In Springer*, 2004, May, pp. 1-19, Berlin, Heidelberg.

[16] 3GPP TS 23.501 System Architecture for the 5G System, Web: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3144. html.

[17] 3GPP TS 23.502 Procedures for the 5G System; Stage 2, Web: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3145. html.

[18] R.M. Kowalski, S.P. Limber, "Psychological, physical, and academic correlates of cyberbullying and traditional bullying", *In Journal of Adolescent Health*, 2013, 53(1), pp.S13-S20.

[19] S.P. Kiriakidis, A. Kavoura, "Cyberbullying: A review of the literature on harassment through the internet and other electronic means", *In Family & community health*, 2010, 33(2), pp.82-93.

[20] K. Dinakar, B. Jones, H. Lieberman, R. Picard, C. Rose, M. Thoman, R. Reichart, "You too?! mixed-initiative lda story matching to help teens in distress", *In Sixth International AAAI Conference on Weblogs and Social Media*, 2012, May.

[21] M. Dadvar, D. Trieschnigg, R. Ordelman, F. de Jong, "Improving cyberbullying detection with user context", *In European Conference on Information Retrieval*, 2013, March, pp. 693-696. Springer, Berlin, Heidelberg.

[22] D. Yin, Z. Xue, L. Hong, B.D. Davison, A. Kontostathis, L. Edwards, "Detection of harassment on web 2.0.", *In Proceedings of the Content Analysis in the WEB*, 2009, 2, pp.1-7.

[23] M. Dadvar, D. Trieschnigg, F. de Jong,"Experts and machines against bullies: A hybrid approach to detect cyberbullies", *In Canadian Conference on Artificial Intelligence*, 2014, May, pp. 275-281. Springer, Cham.

[24] K. Reynolds, A. Kontostathis, L. Edwards, "Using machine learning to detect cyberbullying", *In 10th International Conference on Machine learning and applications and workshops*, 2011, December, Vol. 2, pp. 241-244. IEEE.

[25] S. Park, E.Y. Na, E.M. Kim, "The relationship between online activities, netiquette and cyberbullying", *In Children and youth services review*, 2014, 42, pp. 74-81.

[26] S.V. Sari, "Was it just joke? Cyberbullying perpetrations and their styles of humor. ", *In Computers in Human Behavior*, 2016, 54, pp.555-559.

[27] Q. Huang, V.K. Singh, P.K. Atrey, "Cyber bullying detection using social and textual analysis", *In Proceedings of the 3rd International Workshop on Socially-Aware Multimedia*, 2014, November, pp. 3-6. ACM

[28] Y. Chen, Y. Zhou, S. Zhu, H. Xu, "Detecting offensive language in social media to protect adolescent online safety", *In International Conference on Privacy, Security, Risk and Trust and International Confernece on Social Computing*, 2012, September, pp. 71-80, IEEE.